SpeechTokenizer: Unified Speech Tokenizer for Speech Large Language Models

Year: 2023 | *Citations: 103* | *Authors: Xin Zhang, Dong Zhang, Shimin Li*

Abstract

Current speech large language models build upon discrete speech representations, which can be categorized into semantic tokens and acoustic tokens. However, existing speech tokens are not specifically designed for speech language modeling. To assess the suitability of speech tokens for building speech language models, we established the first benchmark, SLMTokBench. Our results indicate that neither semantic nor acoustic tokens are ideal for this purpose. Therefore, we propose SpeechTokenizer, a unified speech tokenizer for speech large language models. SpeechTokenizer adopts the Encoder-Decoder architecture with residual vector quantization (RVQ). Unifying semantic and acoustic tokens, SpeechTokenizer disentangles different aspects of speech information hierarchically across different RVQ layers. Furthermore, We construct a Unified Speech Language Model (USLM) leveraging SpeechTokenizer. Experiments show that SpeechTokenizer performs comparably to EnCodec in speech reconstruction and demonstrates strong performance on the SLMTokBench benchmark. Also, USLM outperforms VALL-E in zero-shot Text-to-Speech tasks. Code and models are available at https://github.com/ZhangXInFD/SpeechTokenizer/.